ASEE 2014 Zone I Conference, April 3-5, 2014, University of Bridgeport, Bridgpeort, CT, USA.

Encrypted Search & Cluster Formation in Big Data

Gautam Siwach
Department of Computer Science and Engineering
The University of New Haven
West Haven, CT, USA
(gsiwa1@unh.newhaven.edu)

Dr. Amir Esmailpour
Department of Electrical and Computer Engineering
The University of New Haven
West Haven, CT, USA
(Aesmailpour@newhaven.edu)

Abstract - In this paper we investigate the key features of big data as formation of clusters and their interconnections along with their connections to the databases. We focused on the security of big data and the actual orientation of the term towards the presence of different type of data in an encrypted form at cloud interface by providing the raw definitions and real time examples within the technology. Moreover, we propose an approach for identifying the encoding technique in order to perform an expedited search over encrypted text ensuring the security enhancements in big data.

Keywords - Big Data, Cluster, Hadoop, Security

I. INTRODUCTION

In this study we look at the security issues in big data, more specifically to emphasize on an issue of performing the search function when the data is present in cipher form at big data site. For the scope of this research, it is a challenge to perform the search function, as big data comprises a variety of large encrypted data. Hence it is a tough task to fetch the required set of data from trillions of records. Also, big data is an evolving technology and the expectations to be accurate are prominent.

However, the purpose of the study is to present a solution to the existing issue in big data by performing search with the portion of data as plain text over large volume of data in encrypted form that reside in the clusters. The clusters are classified into racks and stacks to store the unformatted data. In order to understand the model we represent the Fig.1 and show the data stored on the stacks connected to form racks which are further consolidated through an advanced 'Hadoop distributed files system' (HDFS) technology framework. Fig.1 also shows the prominent alternatives for datacenter interconnectivity.

The clusters are the backbone in the big data architecture. Securing the technology for privacy and to resolve the issues leading to loss of data are of utter importance. Proper implementation is essential at all stages for the medium to be secure. We emphasize securing the clusters and the network technologies illustrated in the Fig.1 because they seek specific concern in order to substantiate the eradication of vulnerabilities.

Databases works on a very basic concept of input and output of data bits or different volume of packets from disk also called Disk I/O. Data is stored on the disks and the performance of data systems depends on Disk I/O. 'Hadoop' is the next milestone in the field of accessing the stored data. Hadoop technology is introduced as a part of big data technology to operate on the cluster based storage and support Hadoop distributed file system.

A data science in the Fig.1 corresponds to techniques used for fetching the knowledge based on the actual data present in the medium for analyses. It corresponds to Extract, Transform and Load (ETL) tools and techniques to represent the data in a user friendly readable formats charts, figures and diagrams. The Atomicity, Consistency, Isolation and Durability (ACID) properties are marked in the Fig.1 for the guarantee and the reliability of database transactions as part of securing the medium.

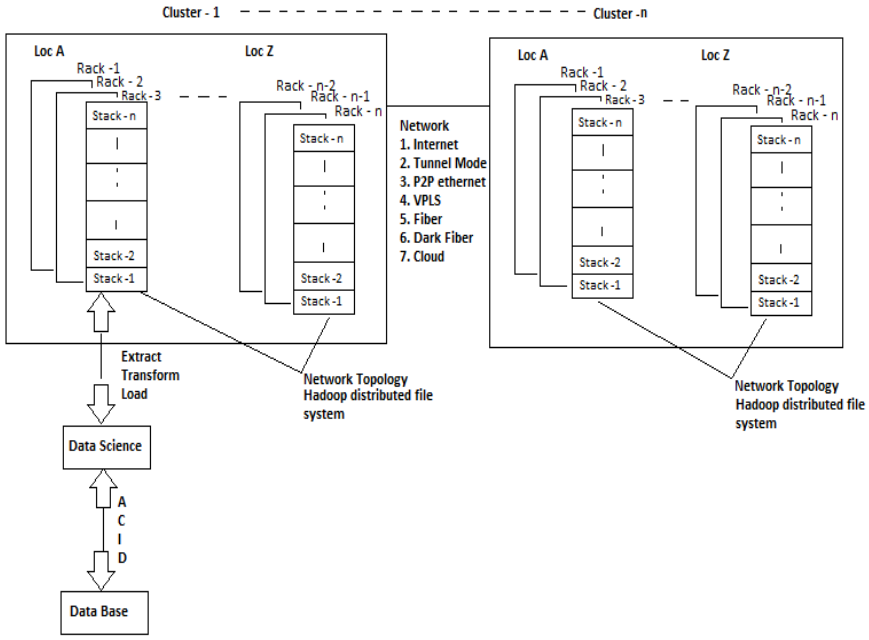

Fig. 1: Cluster concept of the big data technology.

II. BACKGROUND

The security components of big data are implemented at several different points including security at network while data is either in traversal state or at the point of storage with potential for eavesdropping. The data also needs security at the database level. Data protection and security is of high importance and almost everyone's concern that is due to the sensitivity of data ranging from personal, financial details to the data containing national security. Big data holds for the notion, 'the data in this world is exploding or increasing with a very high rate' as shown in Fig.2. A major portion of world's population now has Internet access, and the number of devices is currently estimated in billions. Data is present around us in different form and format that is required for analyses to run a business successfully and competently. If we look at Fig.2 corresponding to the production of digital data we would see that the amount is increasing with almost 2x speed every year as shown in the growth chart for use of big data is depicted followed by year as baseline. The use of data and the need of data analyses are expected to play a major role in providing base to the growth and evolution of yet new analytics and technologies. Big data does not only helps to make data visible to authorized entities but also offers the use of data at higher rate due to the advanced type of distributed file system technology. HDFS technology speeds up the access rate of data by minimizing the time. Use of big data yields more accurate and detailed and global results. The average growth profit rate varies from one organization to the other due to the type of industry.

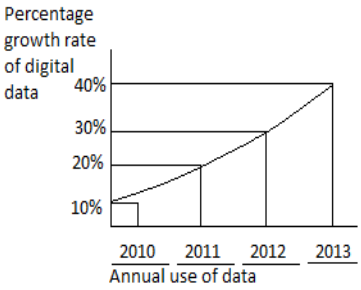

Fig. 2: Growth of big data over past four years.

III. LITERATURE REVIEW

In the research community there has been new movements in big data direction in recent years, several groups around the world work on big data. Jawwad et al (2013) [1] examine the field of cloud computing regarding requirements, expectations, and challenges to provide the related solutions for these challenges. Their study shows the typical classification of intensive cloud with infrastructure and platform in focus and the comparison of NOSQL and B-trees type technologies. For the application specific solutions of data Intensive systems they mentions about the problem of performing search over encrypted data and depicts an instance where a search is to be performed using a portion of data. Later they support the solution to perform search over encrypted data proposed by Wang et al. (2012) by using a rank based key word search scheme.

Alvaro A. Cárdenas et al. (2013) [2] defines the differentiators of traditional and big data thereby emphasizing on volume, variety and velocity of the data. In the paper they investigate security from first generation 'Intrusion detection systems' to third generation 'Big Data in analytics'. In paper, focus is on big data security and the use of cluster Infrastructures that makes it more reliable and available still there remains a scope of improvement. Their analysis leads to the conclusion that big data security is stronger than traditional systems because traditional technologies are limited with long terms and large-scale analytics.

Wang et al. (2012) [3] talks about an important aspect of using more than one data center for a cloud that is obviously argumentative because of several issues associated with it, such as interconnectivity that is choosing an appropriate medium. Their paper also includes finding the perfect hardware and relative software solutions for rank based keyword search scheme.

Adrian Lane (2012) [4] recommends that big data clusters are prone to threats similar to web applications and traditional data warehouses. His paper describes the essential characteristics of big data about most common parallel processing and redundant storage issues. He further questions securing the data and concludes that it is not only confined to big data clusters but the vulnerabilities reside outside the clusters also.

Robert & Yunhong (2009) [5] investigates cloud Infrastructure and types of cloud along with the architectural model and open source cloud. They take Google as a case study for defining the system specific storage. In the paper based on their experiments they infer the necessity of computing the transactions that are supported by Map Reduce or UDF's.

IV. DESIGN AND PROPOSED SOLUTION

In order to design the solution for the problem of performing search over encrypted data where users are not well aware of encrypted data and searching is done over all files repetitively. We propose a design and suggest the use of a unique key to be sent to the user for performing the search. The key acts as a part of private key for user to identify the type of encryption data technique used only to convert the portion of plain text to its corresponding cipher text. Later the encrypted text is used for finding similarities while we perform a search function over encrypted data present in the Data Lake.

We propose a solution by the diagrammatic representation of the concept in big data encryption process. The proposed integrated search concept uses a unique set of key elements tagged with particular form of encryption. To realize the type of encryption technique used, in order to obtain ciphered data. We have implemented the test case scenarios, and successfully tested several use cases to produce an encrypted data, without losing the ACID properties, which ensures the safety of data on board. The encrypted data is present in the Data Lake or in data center, which is connected to the cloud. Fig.3 shows that a client performs search function without using the unique key and gets no results as a match. When the Integrated search function is called using a key then the portion of plain text is converted to corresponding cipher text. Initially the obtained portion of cipher text is matched to the cipher data present in the Data Lake. Upon executing the search function that returns an exact match of data, the obtained encrypted data is retrieved completely and is decrypted to meet the user needs.

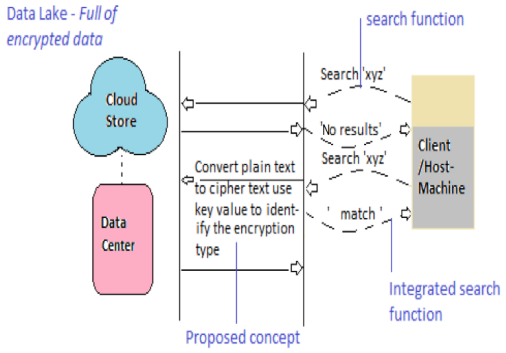

Fig.3: Proposed solution and the concept to perform search over encrypted data.

In the proposed solution the portion of plain text is converted to cipher text using a key. Fig.3 shows the text is present in encrypted form at the cloud of big data integrated to Data Center. To implement, the search function is integrated with proposed solution in a single function. The key is used to identify the encoding type. The integrated search function converts a portion of plain text to the corresponding cipher text present at cloud by identifying the appropriate encoding technique of the cipher text. Once the portion of plain text is converted to cipher text, it will relate to the cipher text present in the cloud and is matched to its corresponding type.

Moreover, we show the encryption process to generate a unique code, which shall correspond to the ciphering technique. Thus it could be used in order to relate to the searched content over the text. A streamlined search of plain text after obtaining its corresponding cipher text can serve as a key to perform the fastest of search. This could allow an intruder to keep away from the data, as there is an option of only converting the plain text to cipher text with the help of unique key element in order to match it to the encoding technique. Hence the entire original plain text used as a reference for search could be converted to cipher text and match could be performed. We are in the initial implementation stage of the solution in Matlab.

V. CONCLUSION

Finally, We have explained the cluster formation and an overview of its architecture. We detailed the consolidated scheme of technological procedures to perform Data Sciences inside Big Data. In addition, we have proposed a solution and drafted the design to implement it in order to perform an integrated search function on the data present

on Data Lake in an encrypted format. The proposed design strictly follows the policies and is set to modularize and test the function before Integration. The simulation shall be tested in order to maintain the security standards and ensure the Quality of service (QOS) on Matlab Environment towards the scope of research. Thereby, increasing the overall performance of the system and initiating new science for futuristic approaches of performing search. The strategy is an effort to identify and access data sets easily and in a rapid manner. The study meets its goal of making a high level design and understanding cluster formation from stacks. The concept is based on the integrated search process inside the proposed solution that continues to provide accurate results. Later the function yields the exact and correct matching results irrespective of the type of encryption technique and form of digital data accession.

REFERENCES

[1] Jawwad Shamsi, Muhammad Ali Khojaye, Mohammad Ali Qasmi, "Data-Intensive Cloud Computing: Requirements, Expectations, Challenges, and Solutions": Journal of Grid Computing, Volume 11 Issue 2, June 2013.Pages 281-310.

[2] Alvaro A. Cárdenas, Pratyusa K. Manadhata , Sree Rajan, "Big Data Analytics for Security Intelligence" : https://cloudsecurityalliance.org/download/big-data-analytics-for-security-intelligence/ Big Data Working Group © 2013 Cloud Security Alliance,September.2013

[3] Cong Wang, Ning Cao, Jin Li, Kui Ren, and Wenjing Lou: "Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data", IEEE Transactions on parallel and distributed systems (vol. 23 no. 8), Aug. 2012.

[4] Adrian Lane: "Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments": https://securosis.com/assets/library/reports/SecuringBigData_FINAL.pdf October. 2012.

[5] Robert L. Grossman, Yunhong Gu: "On the varieties of clouds for data intensive computing". In IEEE data engineering: Bulletin of the IEEE Computer Society Technical Committee on Data Engineering: 2009.